MINERVA– An Efficient Risk-Limiting Ballot Polling Audit

Filip Zagórski, *Wroclaw University of Science and Technology;* Grant McClearn and Sarah Morin, *The George Washington University;* Neal McBurnett*;* Poorvi L. Vora, *The George Washington University*

https://www.usenix.org/conference/usenixsecurity21/presentation/zagorski

This paper is included in the Proceedings of the 30th USENIX Security Symposium.

August 11–13, 2021

978-1-939133-24-3

MINERVA– An Efficient Risk-Limiting Ballot Polling Audit

Filip Zagórski[1], Grant McClearn[2], Sarah Morin[2], Neal McBurnett , and Poorvi L. Vora[2]

[1]Wroclaw University of Science and Technology
[2]Department of Computer Science, The George Washington University

Abstract

Evidence-based elections aim to produce trustworthy and compelling evidence of the correctness of election outcomes, enabling the detection of problems with high probability. They require a well-curated voter-verified paper trail, compliance audits, and a rigorous tabulation audit of the election outcome, known as a risk-limiting audit (RLA).

This paper focuses on ballot polling RLAs which can require that a very large sample of ballots be drawn. The main ballot polling RLA in use today, BRAVO, is designed for use when single ballots are drawn at random and a decision regarding whether to stop the audit or draw another ballot is taken after each ballot draw. But in practice, ballot polling audits draw many ballots in a single round before determining whether to stop.

Direct application of BRAVO to large rounds results in considerable inefficiency. We present MINERVA, a risk-limiting audit that addresses this problem. When compared to the BRAVO stopping rule being applied at the end of the round, for a first-round with 90% stopping probability, MINERVA halves the number of ballots required across all state margins in the 2020 US Presidential election. When compared to the BRAVO stopping rule being applied after examination of individual ballots, MINERVA reduces the number of ballots by about a quarter. MINERVA requires that round sizes are predetermined; this does not appear to be a drawback for large first rounds which have been typical choices for election officials.

Ballot-polling audits are the leading option in most states. MINERVA significantly reduces the necessary expense for contests with close margins and thus makes adopting RLAs easier. Wider adoption of RLAs is a critical step in increasing public confidence in elections.

MINERVA was used in Ohio's pilot RLA of the primaries in May 2020 in Montgomery County. We provide open-source implementations of MINERVA. The code has been integrated as an option in Arlo, the most widely-used RLA software.

*filip.zagorski@votifica.com, Author was partially supported by Polish National Science Centre contract number DEC-2013/09/D/ST6/03927

†poorvi@gwu.edu

1 Introduction

The goal of the Help America Vote Act (HAVA) passed by the United States Congress in 2002 was to improve voting systems, but it led to the large-scale deployment of insecure electronic voting systems [2–4]. Additionally, no matter how well-developed a voting system, it is not possible to be certain of its output in all instances. Best practices hence should include *software independent* [16] voting systems (those in which an undetected error in the software does not result in an undetectable error in the election outcome) and *evidence-based elections* [21], in which the voting system provides not only the tally, but also evidence that the outcome is correct, and in which the evidence is examined by the public to determine whether the outcome is correct.

The more recent trend is a return to paper-based systems, albeit with ongoing voter verification weaknesses in some quarters [1]. Voter-verified paper records, combined with secure curation of the paper trail and compliance and voter registration audits provide an independent record of voter intent and convert almost any voting system into a system that is evidence-based. One need not perform a full manual recount of the independent record of voter intent to verify the election outcome. Risk Limiting Audits (RLA), as described by Lindeman and Stark [8], provide a rigorous approach to confirming the election outcome through the sampling of a subset of the ballots. A report from the National Academy of Sciences [13] and the Voluntary Voting Systems Guidelines (VVSG, version 2.0) [22] strongly support the use of RLAs.

Significant effort has been invested by policy advocacy organizations—such as Verified Voting, the Brennan Center, Common Cause, Democracy Fund and others—to educate election officials about RLAs and help them carry out pilots and statutory audits, as well as develop legislation. Non-profit VotingWorks has often been a partner, providing both open-source audit software (Arlo) and training in its use. As a consequence of these efforts, three states (Colorado, Rhode Island and Virginia) have RLAs in statute; four have a statutory pilot program (Georgia, Indiana, Kentucky, and Nevada, where RLAs will be a requirement in 2022); four allow RLAs to satisfy a more general audit requirement (California, Ohio, Oregon and Washington); and two have an administrative pilot program (Michigan and New Jersey).

Of the different types of RLAs, the ballot polling RLA requires only minimal information—the tally and an independent ballot manifest describing the organization of ballot storage. For close contests, however, a large number of ballots need to be drawn at random, requiring considerable effort on the part of election officials. These are early times in the adoption of RLAs, and difficulties could scuttle these efforts. It is hence in the interests of election security that unnecessary inefficiencies be rectified.

We describe the MINERVA ballot polling RLA, which considerably reduces the workload when compared to BRAVO, the most common ballot polling audit. We show that MINERVA is an RLA if the round schedule is pre-determined.

1.1 Election Tabulation Audits

An election tabulation audit may be viewed as a binary hypothesis test, where the null hypothesis is that the election outcome is incorrect, and the alternative hypothesis is that it is correct. An audit is defined as *risk-limiting* to *risk limit* α if its Type I error is at most α, whatever the (unknown) true underlying vote distribution is. That is, given that the tabulated outcome is incorrect, the probability that the audit does not recognize it as such is at most α.

The most popular example of election tabulation ballot polling audits is the BRAVO [9] audit, which is a *most efficient* audit when the audit software is queried (regarding whether to stop the audit or draw more ballots) after each ballot draw, and the stopping condition is satisfied exactly when the audit is stopped. The term *most efficient* refers here, as elsewhere, to an audit requiring the smallest expected number of ballots if the election is drawn from the assumed prior. The expectation is taken over the randomness of the ballot draws.

In real election audits, multiple ballots are drawn in a round before a decision is taken. This paper shows that BRAVO is not a most efficient test in this case and proposes the more efficient risk-limiting test MINERVA, demonstrating significant decreases in first-round sizes, or lower Type I errors for the same number of ballots. For example, consider a ballot polling RLA of the 2020 Presidential contest in the state of Michigan. Election officials generally prefer a high probability of being done in a single round. If the audit applied the BRAVO rule at the end of the round, the first round size for a 90% probability of completion is 18,161 (measured in the expected number of distinct ballots). The corresponding round size for MINERVA is 8,807. Additionally, if the audit is not completed in the first round, the MINERVA Type 1 error measure is always smaller than the BRAVO measure for the same sample.

To reduce the expected BRAVO sample size, auditors can apply some additional bookkeeping effort so the BRAVO rule can be applied to each ballot in random selection order after the ballots are drawn in a round. With this approach, the first round size for a 90% probability of completion with BRAVO would drop to 12,293, still almost 40% larger than the MINERVA round size.

1.2 The Problem

We refer to audits where decisions are taken after each ballot draw as *ballot-by-ballot* or *B2* audits. The general audit, however, is a *round-by-round* or *R2* audit where, in the j^{th} round, some ballots are drawn, after which a decision is taken regarding whether to (a) stop the audit and declare the election outcome correct, (b) stop the audit and go to a full manual recount, or (c) draw the $(j+1)^{th}$ round. A B2 audit is a special case of the R2 audit, when a single ballot is drawn in each round.

There are two ways to apply B2 audit rules to an R2 audit. Let n be the number of ballots drawn at any time. Let n_j be the total number of ballots drawn after the j^{th} round, of which k_j are for the reported winner. Hereafter, when we refer to the "winner" we typically mean the "reported winner"; we refer to the "true winner" when necessary.

- *End-of-round:* In this application, the B2 stopping rule for k_j winner ballots in a sample of n_j ballots determines whether the audit will stop.
- *Selection-ordered-ballots:* In this application, the interpretation of each ballot is associated with the ballot id, so the B2 stopping condition can be tested $\forall n \leq n_j$. The audit stops if the B2 condition is satisfied for any value of $n \leq n_j$.

Selection-ordered-ballots is generally more efficient than *end-of-round* as a means of applying B2 rules to R2 audits, but requires the significant additional effort of preserving enough information to be able to recreate the subtotals of winner ballots in selection order. *End-of-round* relies only on the tallies and does not require selection order. As our paper shows, neither is a most efficient R2 stopping rule.

1.3 Our Contributions

Our contributions are as follows:

1) We derive analytical expressions for the risk and probability of stopping, given the history of rounds and the election margin for the BRAVO stopping rule. For rounds drawing single ballots each, we verify that our expressions predict the stopping percentiles of BRAVO simulations as reported by Lindeman *et al.* [9, Table 1]. The code for computing these expressions is available as a MATLAB library, released as open-source under the MIT License [23].

2) We present the MINERVA audit (or stopping rule) and prove that it is risk-limiting and at least as efficient as the corresponding *end-of-round* BRAVO stopping rule, and, relatedly,

that MINERVA provides a risk limit that is never larger than the one for *end-of-round* BRAVO.

Note that our current proof for the MINERVA properties assumes that the audit uses a predetermined schedule of round sizes, independent of what samples are drawn in earlier round sizes. Because BRAVO is designed for use with rounds of size one, a BRAVO audit need not pre-commit to round sizes. Because it appears that election officials prefer not to go back for another round, and our efficiency gains are considerable for high stopping probabilities in the first round, MINERVA is useful in spite of this constraint. For example, one could choose the second round size for MINERVA to correspond to the total number of ballots to be drawn by BRAVO in its first round, and one would see that MINERVA is virtually certain to stop by then if the required stopping probability in the first round is large enough.

3) We provide experimental results and software to support the use of MINERVA:

- To illustrate the efficiency improvements, we compute (without simulations, using the derived analytical expressions), for each state in the 2020 US Presidential election, risk limit $\alpha = 0.1$ and a stopping probability of 0.9, first round sizes for *end-of-round* BRAVO and MINERVA. We find that *end-of-round* BRAVO requires about twice the number of ballots, across all margins.

- We compute first round sizes for *selection-ordered-ballots* BRAVO and find that it requires about $25-39.5\%$ more ballots for the data of the 2020 US Presidential election, with the improvement due to MINERVA being better for smaller margins. Thus MINERVA is more efficient than *selection-ordered-ballots* BRAVO and does not require the additional bookkeeping of recording selection ballot order.

- Our code for the audits is available as MATLAB and Python libraries [12,23,28]. All code is released as open source under the MIT license. The Python code for the MINERVA audit was used for an RLA pilot in Montgomery County, Ohio in May 2020. The Python library has been integrated as an option into Arlo, the most popular election audit software that has been used to run a large number of RLAs [25].

Advocacy groups continue to work towards policies supporting or requiring RLAs, and efficiency improvements will impact their adoption and progress towards evidence-based elections.

For those implementing audits, we note that, depending on the margin and the voting technology used, ballot comparison or batch comparison audits could be more desirable RLAs. One may also consider combinations of ballot comparison and ballot polling audits, such as described in [14].

4) The class of R2 stopping rules is a class of B2 rules when round size is one. Of theoretical interest, we prove that B2 MINERVA (round size one) has the same stopping rule as B2 BRAVO. We do not claim that MINERVA is a most efficient R2 audit; the problem of finding the most efficient R2 audits is open.

1.4 Organization

Section 2 presents the model and related work. Section 3 motivates the problem with an example demonstrating that the application of B2 rules to an R2 audit results in inefficiencies. Section 4 introduces the MINERVA audit with examples and provides insight into why the audits are risk-limiting and more efficient than either R2 application of B2 BRAVO. Section 5 presents rigorous claims of MINERVA's risk-limiting and efficiency properties. Section 6 presents applications of our results, and Section 7 describes the use of MINERVA in a pilot RLA of the primaries in Montgomery County, Ohio in May 2020. Section 8 concludes. Some proofs and some more experimental results are in the Appendix.

2 Background

There are three main categories of risk-limiting audits. Each requires that the voting system produce a tally, and that administrators provide a ballot manifest, independent of the voting system, which details the physical storage organization of the ballots so that specific ballots may be identified (for example, "the fifth ballot in Batch 100"). Each has some distinguishing traits:

1. Ballot polling audits: Ballots are drawn at random and a decision of whether to stop the audit or not is based on the sample drawn. No additional data is required.

2. Batch comparison audits: The ballots are organized in batches, such that each ballot belongs to exactly one batch. Batches are chosen at random and the manual tally of all ballots in a chosen batch is compared to the batch tally announced by the voting system. The results of the comparisons determine whether to stop the audit or draw another batch/batches. This type of audit requires the voting system to provide tallies for each batch. It thus requires more granular detail about the tally than does a ballot polling audit.

3. Ballot comparison audits: Individual ballots are chosen at random and the physical ballot is compared to the cast vote record (CVR), which is the electronic representation of the ballot choices as interpreted by the voting system. The collection of CVRs uniquely determines the tally. The results of the comparisons provide the basis for the stopping decision. This type of audit requires the voting system to provide a CVR for each ballot. It thus requires more granular detail about the tally than do the ballot polling audit and the batch comparison audit. It is

the most efficient of all three audits in terms of number of ballots drawn, but many voting systems either don't produce CVRs, or make them too difficult to match up with the corresponding paper ballots.

The comparison audits may provide greater satisfaction to election officials and the public because each comparison provides a measure of how close the system came to achieving the goal of a correct election. But batch comparison audits can require more setup and work than a ballot polling audit, especially when batches are large. Ballot comparison audits are not always feasible. For these reasons, ballot polling audits are the most common and have been used in a number of state pilots (California, Georgia, Indiana, Michigan, Ohio, Pennsylvania and elsewhere). The main challenge in implementing ballot polling audits is the very large number of ballots that need to be drawn in close contests; $\Theta(\frac{1}{m^2})$ for margin m. Note that the multiplying constant for the audit does matter in this scenario; a factor of two improvement, for example, can reduce by nearly half the number of person days required to complete a large audit. This is particularly useful when the number of ballots to be drawn is in the tens of thousands, such as would have been the case for ballot polling RLAs with risk limit 0.1 in Georgia, Arizona, Wisconsin, Pennsylvania, North Carolina, Nevada, Michigan and Florida in the 2020 Presidential contest (see Table 1).

2.1 The Model

We consider a plurality contest and assume ballots are drawn with replacement. We assume all ballots have a vote for either the winner or the loser; because ballots are sampled with replacement, our argument is easily extended to contests with multiple candidates and invalid ballots (as for BRAVO, for example, see [8]). We denote by w the true winner, w_a the announced winner, ℓ_a the announced loser and p the announced fractional tally for w_a (typically based on preliminary, uncertified results).

A polling audit will estimate whether w_a is the true winner. We denote by n_j the total number of ballots drawn at the end of the j^{th} round, and by k_j the corresponding total number of ballots for the winner. Hence the number of new ballots drawn in round j is $n_j - n_{j-1}$, and the number of new votes for the winner drawn in round j is $k_j - k_{j-1}$. If necessary, one may assume that $n_0, k_0 = 0$. We often refer to $[n_1, n_2, \ldots, n_j, \ldots]$ as the *round schedule*. A B2 audit is an R2 audit with round size $n_j = j$. That is, the round schedule of a B2 audit is $[1, 2, \ldots, j, \ldots]$.

The total number of ballots drawn at any time during the audit is denoted n (if the number of rounds drawn so far is j, $n = n_j$). The random variable representing the number of ballots drawn so far for the winner is represented by K. We use k^*, k_* and $\tilde{k}$ to represent specific numbers of winner ballots as well.

The entire sample drawn up to the j^{th} round, in sequence, forms the *signal* or the *observation*; the corresponding random variable is denoted X_j, the specific value x_j. The entire sample drawn so far is denoted X, its specific value x. We do not *a priori* assume a last round for the audit. The audit stops when it satisfies the stopping condition.

We model the audit as a binary hypothesis test:

Null hypothesis H_0: The election outcome is the closest possible incorrect outcome: $w \neq w_a$ and the fractional vote count for w_a is $\frac{1}{2}$. In particular, if the total number of valid votes is even, the election is a tie. If the total number of valid votes is odd, the margin is one in favor of ℓ_a. In this case, we assume that the number of valid votes is large enough that the fractional vote count is sufficiently close to $\frac{1}{2}$. Henceforth, we will refer to both cases as being represented by a fractional vote count of $\frac{1}{2}$.

Alternate hypothesis H_a: The election outcome is correct: $w = w_a$ and the fractional vote count is as announced.

After each round the test $\mathcal{A}$ takes X as input and outputs one of the following:

- *Correct:* The test estimates that $w = w_a$ and the audit should stop.
- *Incorrect:* The test estimates that $w \neq w_a$. We stop drawing votes and proceed to perform a complete hand count to determine w.
- *Undetermined (draw more samples):* We need to draw more ballots to improve the estimate.

When the audit stops, it can make one of two kinds of errors:

1. *Miss:* A *miss* occurs when $w \neq w_a$ but the audit misses this, and outputs *Correct*. We denote by P_M the probability of a miss:

$$P_M = Pr[\mathcal{A}(X) = \textit{Correct} \mid H_0]$$

P_M is the *risk* in risk limiting audits and the Type I error of the test.

2. *Unnecessary Hand Count:* Similarly, if $w = w_a$, but the audit estimates that a hand count must follow, the hand count is unnecessary. We denote the probability of an *unnecessary hand count* by P_U:

$$P_U = Pr[\mathcal{A}(X) = \textit{Incorrect} \mid H_a]$$

P_U is the Type II error.

Like the BRAVO audit, this paper focuses on tests with $P_U = 0$. The risk, on the other hand, is an important (generally) non-zero value characterizing the quality of the audit.

2.2 Related Work

A *risk-limiting audit (RLA)* with *risk limit* α—as described by, for example, Lindeman and Stark [8]—is one for which the risk is smaller than α for all possible (unknown) true tallies in the election. For convenience when we compare audits, we refer to this audit as an α-*RLA*.

Definition 1 (Risk Limiting Audit (α-RLA))**.** *An audit $\mathcal{A}$ is a Risk Limiting Audit with risk limit α iff*

$$P[\mathcal{A}(X) = Correct \mid H_0] \leq \alpha$$

There are many audits that would satisfy the α-*RLA* criterion, and not all would be desirable. For example, the constant audit which always outputs *Incorrect* always requires a hand count and is risk-limiting with $P_M = 0 < \alpha, \forall \alpha, \forall p$. However, $P_U = 1$, and the audit examines all votes each time; this is undesirable.

An example of an α-*RLA* with $P_U = 0$ and drawing fewer ballots is the B2 BRAVO audit [9] which specifies round size increments of one.

We use the following notation:

$$\sigma(k,p,n) \triangleq \frac{p^k(1-p)^{n-k}}{(\frac{1}{2})^n} \tag{1}$$

Definition 2 (BRAVO)**.** *An audit $\mathcal{A}$ is the B2 (α, p)-BRAVO audit iff the following stopping condition is tested at each ballot draw. If the sample X is of size n and has k ballots for the winner,*

$$\mathcal{A}(S) = \begin{cases} Correct & \sigma(k,p,n) \geq \frac{1}{\alpha} \\ Undetermined & else \end{cases} \tag{2}$$

Its p-value is $\sigma(k,p,n)^{-1}$.

$\sigma(k,p,n)$ is the *likelihood ratio* of the drawn sequence X. The B2 (α, p)-BRAVO audit is an *SPRT* [26] with:

H_0, the null hypothesis: the election is a tie

H_a, the alternate hypothesis: the fractional tally for the winner is p.

Implicit in Definition 2 is the point that a sequence X is tested only if it has not previously satisfied the test. If $\mathcal{A}(X_*) = Correct$ for some sequence X_*, all extensions X_*^+ of X_* are defined as having passed the test. Determining the stopping condition by evaluating $\mathcal{A}(X_*^+)$ does not satisfy the assumptions of the test, and the properties of the test do not necessarily apply. As we shall see in Section 3, this is relevant to *end-of-round* BRAVO. In fact, it is relevant to *end-of-round* applications of any B2 audit that is an *SPRT*.

B2 BRAVO is a most efficient test given the hypotheses (if, in each instance that the stopping condition is satisfied, it is satisfied exactly). Vora shows that B2 (α, p)-BRAVO is an α-*RLA* because it assumes a tie for H_0, which is the wrong election outcome that is hardest to distinguish from the announced one, and hence defines the worst-case risk [24].

Other approaches, such as Rivest's CLIP Audit [15], improve on B2 BRAVO's efficiency subject to certain constraints (namely, of β as defined in [15]). More creative approaches, such as the k-cut method, attempt to reduce the effort made by election officials in a ballot polling audit [18].

An early prototype of MINERVA mirrored the explicit risk allocation found in Stark's Conservative Statistical Post-Election Audits [20]: before ballots are examined for the audit, a list of increasing rounds $(n_1, n_2, ..., n_j)$, and a list of corresponding risks $(\alpha_1, \alpha_2, ..., \alpha_j)$ are generated. The early prototype solved this problem by exactly computing the risk and probability distributions (using the convolution as described in a later section). This led to a fundamental improvement of MINERVA over BRAVO.

There is a line of work on *group sequential testing* [5–7, 27] but all results that we were able to find begin with the assumption of a normal distribution and cannot be directly applied to the considered scenario of auditing elections.

3 BRAVO Theory vs Practice

In this section we use an example to illustrate the problems of using B2 rules for an R2 audit.

The B2 (α, p)-BRAVO audit, Definition 2, is the following ratio test (inequality (2)) performed after each draw: $\sigma(k,p,n) = \frac{p^k(1-p)^{n-k}}{(\frac{1}{2})^n} \geq \frac{1}{\alpha}$.

Because $p > 1-p$ and the denominator above does not depend on k, $\sigma(k,p,n)$ is monotone increasing with k. There is hence a minimum value of k for which the B2 (α, p)-BRAVO stopping condition is satisfied. That is, $\exists\, k_{min}(\text{BRAVO}, n, p, \alpha)$ such that the stopping condition of Definition 2, inequality (2), is:

$$\mathcal{A}(S) = Correct \Leftrightarrow k \geq k_{min}(\text{BRAVO}, n, p, \alpha)$$

In fact it is easy to see that $k_{min}(\text{BRAVO}, n, p, \alpha)$ is a discretized straight line as a function of n, with slope and intercept determined by p and α (see, for example, [26]).

$$k_{min}(\text{BRAVO}, n, p, \alpha) = \lceil m(\text{BRAVO}, p, \alpha) \cdot n + c(\text{BRAVO}, p, \alpha) \rceil \tag{3}$$

where

$$m(\text{BRAVO}, p, \alpha) = \frac{\log \frac{\frac{1}{2}}{1-p}}{\log \frac{p}{1-p}}$$

$$c(\text{BRAVO}, p, \alpha) = -\frac{\log \alpha}{\log \frac{p}{1-p}}$$

We drop one or more arguments of k_{min}, c or m when they are obvious.

Example 1 (B2 BRAVO vs R2 BRAVO). Let $\alpha = 0.1$ and $p = 0.75$, we get, from equation (3):

$$k_{min}(\text{BRAVO}, n, 0.75, 0.1) \approx \lceil 0.6309n + 2.0959 \rceil \quad (4)$$

Consider ballots drawn in rounds of size $20, 40, 60, \ldots$ and the BRAVO condition being tested:

- *End-of-Round*, which requires a record simply of the tally of the sample polled.
- *Selection-ordered-ballots*, requires a record of the vote on each ballot polled, in selection order.

Note that the stopping condition is always the BRAVO stopping condition; the variation is in when it is checked.

Figure 1 is a plot of $k_{min}(\text{BRAVO}, n, 0.75, 0.1)$ as a function of round size. It also shows the results of the tests above, performed on an example sequence.

- For a hypothetical sequence, *selection-ordered-ballots* BRAVO checks the stopping condition at the blue squares till the stopping condition is satisfied, and the audit stops. It has information about the number of ballots for the winner and the total number of ballots drawn at each ballot draw.
- If the same sequence were to go through an *end-of-round* BRAVO audit, the stopping condition would be checked only at the end of the round, denoted in the figure by black crosses. The audit only has information on vote tallies at the end of the round.

We see that the stopping condition is satisfied during the second round, at $n = 22$, but that it is no longer satisfied when it is tested at the end of that round, at $n = 40$, or the following round, $n = 60$. It is satisfied at the end of the fourth round, $n = 80$, which is the number of ballots drawn in an *end-of-round* BRAVO audit. Thus:

- B2 BRAVO ends at $n = 22$, and 22 ballots are drawn.
- *End-of-round* BRAVO ends at $n = 80$ and 80 ballots are drawn.
- *Selection-ordered-ballots* BRAVO ends at $n = 22$, and 40 ballots are drawn.

The instance of *selection-ordered-ballots* BRAVO in our example would stop at the end of the second round after 40 ballots are drawn, but the information in ballots 23-40 would be discarded. It ought to be possible to use this information, obtained at some cost, to better estimate the correctness of the election outcome. (Imagine telling election officials and the public that the p-value of the draw was small enough earlier, that it is not any more, and the math allows us to use the earlier value because if the election outcome is incorrect, it is accounted for in the risk limit). We need not be limited by the B2 BRAVO rules which begin with a large disadvantage when used for R2 audits, as they do not take into account that the ballots are drawn in rounds.

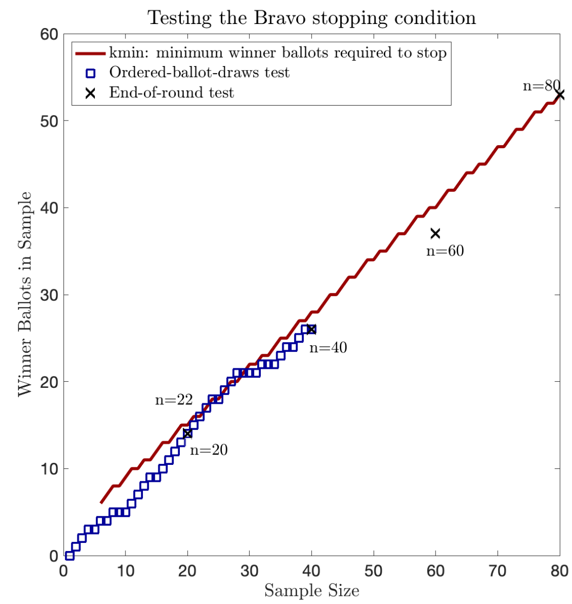

Figure 1: Using BRAVO for a round-by-round audit with $p = 0.75$, $\alpha = 0.1$ and round size = 20.

4 MINERVA

In this section, we use an example to illustrate the workings of a proposed new R2 audit MINERVA. In later sections, we prove MINERVA is risk-limiting—if round sizes are predetermined—and at least as efficient as *end-of-round* BRAVO.

4.1 End-of-round BRAVO

Example 2 (*End-of-Round* $(0.1, 0.75)$-BRAVO). We consider the *end-of-round* $(0.1, 0.75)$-BRAVO audit as in the previous section. Denote by n_1 the number of ballots drawn in the first round and let $n_1 = 50$. Let K_1 be the number of votes for the winner, then K_1 lies between 0 and $n_1 = 50$. Figure 2 shows the probability distributions of K_1 for the two hypotheses:

H_a: the election is as announced, with $p = 0.75$ (blue solid curve), and

H_0: the election is a tie (red dashed curve).

We will continue to refer to Figure 2 in the following examples and sections, when we will address the shaded areas.

4.2 An Introduction to MINERVA

We propose the MINERVA audit, which uses the tails of the probability distribution functions to define the stopping con-

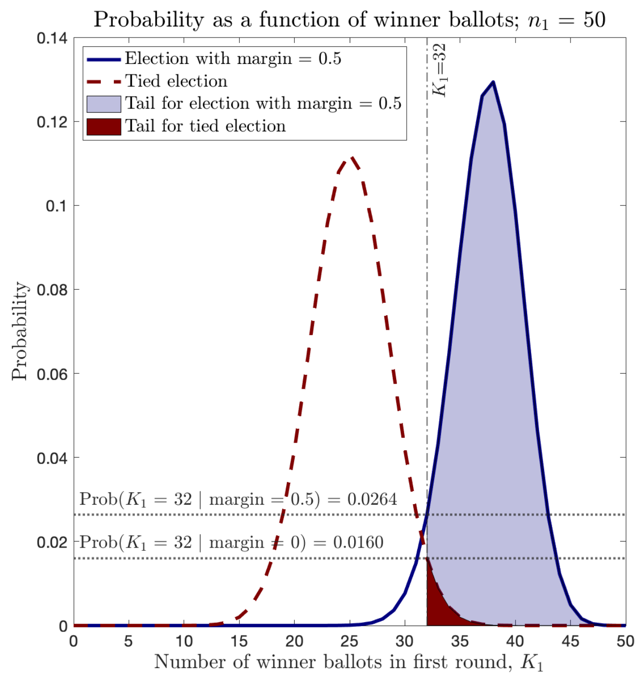

Figure 2: *Probability Distribution of Winner Votes for a fractional vote count of* $p = 0.75$ *for the winner and a sample size of* $n_1 = 50$*: First Round.* The solid blue line represents the distribution if the election tally is as announced. The dashed red line represents the distribution if the election is actually a tie, which—among all elections not won by the announced winner—is the hardest to distinguish from an election actually won by the announced winner. If the number of ballots drawn for the winner were $k_1 = 32$, the horizontal dotted lines mark the values where the vertical line $k_1 = 32$ crosses the two distributions. These are the values needed to compute the BRAVO ratio, σ, and correspond to $Pr(K_1 = 32 \mid margin = 0.25)$, which is 0.0264, and $Pr(K_1 = 32 \mid margin = 0)$, which is 0.0160. The BRAVO ratio is then $\sigma(32, 0.75, 50) = \frac{0.0264}{0.0160} \ngeq \frac{1}{\alpha} = 10$ and the sample does not pass the *end-of-round* BRAVO audit in the first round. Recall that the B2 BRAVO p-value is the reciprocal of the above probability ratio. In this example, it is $\approx 0.6061 > \alpha = 0.1$, and the maximum risk is larger than the risk limit. This is consistent with the fact that (see Example 1, Section 3) $32 < k_{min}(\text{BRAVO}, 50, 0.75, 0.1) = 34$.

dition. Here we provide an informal description of the MINERVA audit.

We denote:

$$\tau_1(k, p, n_1) = \frac{Pr[K_1 \geq k \mid H_a, n_1]}{Pr[K_1 \geq k \mid H_0, n_1]} \tag{5}$$

Example 3 (The MINERVA Audit)**.** For the parameters of Example 2, $\alpha = 0.1$, $p = 0.75$, $n_1 = 50$ and $k_1 = 32$, we describe the MINERVA stopping condition, a comparison test of the ratio of the tails of the distributions:

$\tau_1(32, p, n_1) \geq \frac{1}{\alpha}$.

Compare this to the stopping condition for BRAVO, inequality (2).

Note that $Pr[K_1 \geq 32 \mid H_a]$ is the stopping probability for round 1 (the probability that the audit will stop in round 1 given H_a) associated with deciding to stop at $K_1 = 32$—and not at smaller values. It is the tail of the solid blue curve, the translucent blue area in Figure 2. Similarly, $Pr[K_1 = 32 \mid H_0]$ is the associated risk. It is the tail of the red dashed curve denoting the tied election, and shaded red.

For our example, the ratio of the tails of the two curves of Figure 2 is (the values are not denoted in the figure):

$\tau_1(32, 0.75, 50) = \frac{Pr[K_1 \geq 32 | H_a, n_1]}{Pr[K_1 \geq 32 | H_0, n_1]} \approx 29.89 > \frac{1}{\alpha} = 10$.

And the sample passes the MINERVA audit.

We see below that the MINERVA ratio is larger than the BRAVO ratio for various values of K_1 in our example. We show more rigorously later that the MINERVA ratio is always no smaller than the BRAVO ratio, and hence that MINERVA is always at least as efficient as *end-of-round* BRAVO. We also show later that the MINERVA test is risk-limiting.

Example 4 (BRAVO vs. MINERVA Ratios)**.** For the parameters of Examples 2 and 3: $p = 0.75$, $\alpha = 0.1$ and $n_1 = 50$, Figure 3 presents the likelihood ratio for *end-of-round* BRAVO (green solid line), $\sigma(k_1, 0.75, 50)$, and the tail ratio for MINERVA (orange dashed line), $\tau_1(k_1, 0.75, 50)$, on a log scale. An audit satisfies the stopping condition when its ratio equals or exceeds $\alpha^{-1} = 10$, and we observe in the figure that the MINERVA audit stops before *end-of-round* BRAVO.

This is an example of a more general relationship: any sample satisfying *end-of-round* BRAVO will also satisfy MINERVA. In fact, it will often be the case that the MINERVA condition will be satisfied and that for *end-of-round* BRAVO will not. The reason for MINERVA stopping at smaller values of K_1 is as follows.

The MINERVA ratio, τ_1, at some $K_1 = k_1$ is a weighted average of all the values of $\sigma(K_1, 0.75, 50)$ for $K_1 \geq k_1$. Because $\sigma(K_1, 0.75, 50)$ is an increasing function of K_1, the weighted average is, generally speaking, larger than the value of $\sigma(k_1, 0.75, 50)$, because the larger values of $\sigma(K_1, 0.75, 50)$ "make up" for the smaller ones. It is never smaller than $\sigma(k_1, 0.75, 50)$. When $k_1 = n_1$ is the largest possible number of winner votes, the two ratios will be equal. Equivalently, the MINERVA p-value will always be smaller than the BRAVO one, except when $k_1 = n_1$ is the largest possible number of winner votes, and the p-values are equal. Thus, MINERVA always stops when *end-of-round* BRAVO does, and is at least as efficient.

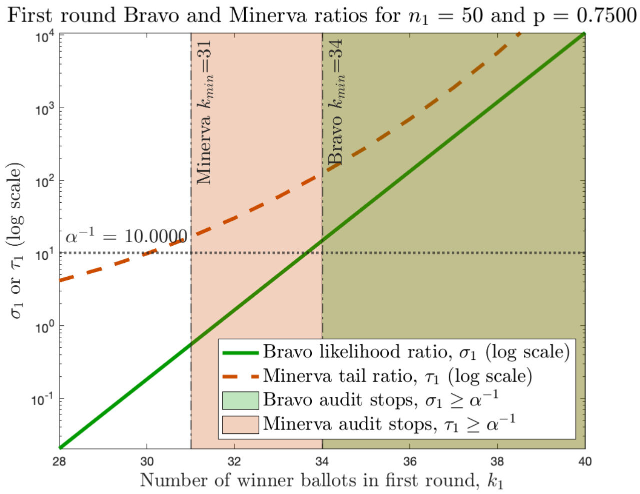

Figure 3: BRAVO *and* MINERVA *comparison tests for* $p = 0.75$ *and* $n_1 = 50$*: First Round.* The figure above presents the BRAVO and MINERVA ratios (σ and τ respectively) as a function of the number of winner ballots for a sample size of $n_1 = 50$. The orange dashed line, representing the MINERVA ratio, is above the solid green line representing the BRAVO ratio. The horizontal line $\alpha^{-1} = 10$ marks the value above which the stopping condition is satisfied; that is, the MINERVA stopping condition is $\tau_1 \geq \alpha^{-1}$ and the BRAVO stopping condition is $\sigma_1 \geq \alpha^{-1}$. We have seen earlier that the k_{min} value for BRAVO for these parameters is 34 (see equation (4), Example 2), and this is consistent with what we see in the figure: *end-of-round* BRAVO stops for $K_1 \geq 34$ (shaded green area) and no smaller values of K_1. On the other hand, we see from this figure that MINERVA stops for $K_1 \geq 31$ (shaded orange area) and is hence more efficient for this round.

4.3 Computing Risks and Stopping Probabilities for Multiple-Round Audits

In this section we describe how probability distributions may be computed in multiple round audits with monotone stopping conditions; that is, audits where the stopping condition is represented through the use of k_{min}. We use examples to demonstrate how the probability distributions may be computed for rounds 2 and above.

Example 5 (Testing the Stopping Condition)**.** Consider an election with $p = 0.75$ and a risk limit of $\alpha = 0.1$. Suppose the first round size is $n_1 = 50$ and the draw results in K_1 ballots for the announced winner. Recall that the k_{min} value for MINERVA for these parameters is 31 (see Figure 3, Example 4) and we assume that $K_1 < k_{min}$. Thus the sample does not pass the MINERVA test.

Now suppose we draw 50 more ballots to get $n_2 = 100$ ballots in all, of which K_2 are for the winner. We will need to compute the probability distribution on K_2 to determine the ratio of the tails for the MINERVA stopping condition.

Note that the probability distribution of K_2 is *not* the binomial distribution for a sample size of 100. In fact, if the audit did not stop in the first round, $K_1 < 31 = k_{min}$ and $K_2 \leq K_1 + 50$, which means that $K_2 < 81$ (even if all 50 ballots in the second round are for the announced winner).

If the audit continues, the maximum number of ballots before new ones are drawn is 30. The probability distributions before the new sample is drawn are as shown in Figure 4, and may be denoted as:
$f(K_1 \mid H_a) = Pr[K_1 = k_1 \wedge (\mathcal{A}_\mathcal{M}(X_1) \neq Correct)|H_a]$ and
$f(K_1 \mid H_0) = Pr[K_1 = k_1 \wedge (\mathcal{A}_\mathcal{M}(X_1) \neq Correct)|H_0]$.
where $\mathcal{A}_M$ denotes the MINERVA audit for the given parameters.

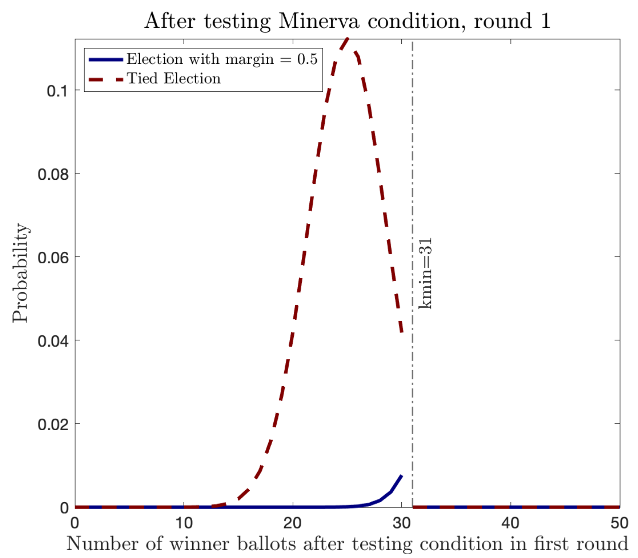

Figure 4: *Probability Distribution of Winner Ballots for* MINERVA*:* $p = 0.75$*,* $n_1 = 50$*: After Testing the Stopping Condition for the First Round.* Going into the second round, the values of the probability distributions for $K_1 < k_{min}$ are unchanged, and the probabilities for $K_1 \geq k_{min}$ are zero, because the audit would have stopped if $K_1 \geq k_{min}$.

The “discarded” tails, in both cases, represent the probabilities that the audit stops. When this is conditional on H_a, we refer to it as the stopping probability of the round (S_1), large values are good. When it is conditional on H_0, it is the worst-case risk corresponding to the round (R_1), large values are bad. Recall that our stopping condition bounds the worst-case risk for the round to be no larger than a fraction α of the stopping probability.

Using the above probability distributions, we can now compute the distribution of ballots for the announced winner in the sample of size 100, which we obtain after drawing 50 more ballots.

Example 6 (Second Round Distribution)**.** Continuing with Example 5, we consider an election with $p = 0.75$, risk limit $\alpha = 0.1$ and round sizes $n_1 = 50, n_2 = 100$. We wish to compute the probability distribution for K_2, the number of votes for the announced winner after drawing the second round of ballots.

There would be a total of K_2 winner ballots in the sample after the second draw if $D = K_2 - K_1$ winner ballots were drawn among the 50 new ballots drawn in round 2. D could be any value between 0 and 50, and its distribution is the binomial distribution for the draw of size 50.

If we denote the distribution of K_2 as g, it is:

$$g(K_2 = k_2 \mid H) =$$

$$\sum_{k_1=max\{0,k_2-50\}}^{min\{k_{min}-1,k_2\}} f(K_1 = k_1 \mid H) \cdot Bin(k_2 - k_1, 50, H)$$

where $Bin(j,n,H)$ is the probability of drawing j votes for the announced winner in a sample of size n, when the fractional vote for the announced winner is $\frac{1}{2}$ for $H = H_0$ and p for $H = H_a$.

The above expression results in a function g_H, obtained by an operation known as the convolution of the two functions, and is denoted:

$$g_H = f_H \circledast Bin_{H,50}$$

where $\circledast$ represents the convolution operator and H the hypothesis. The convolution of two functions can be computed efficiently using Fourier Transforms; this result is the convolution theorem.

After drawing the second sample, the probability distributions for MINERVA are as in Figure 5.

In order to compute probability distributions for the next round, we would first compute the value of k_{min} for this round using the tail ratio, then zero the probability distributions for the value of k_{min} and above, and then perform a convolution with the binomial distribution corresponding to the size of the next draw. And so on.

Probability distributions for B2 audits may be computed similarly, with the round schedule: $(1,2,\ldots,i,\ldots)$. We used this approach to compute percentiles for the BRAVO stopping probabilities; see the Appendix for the results.

4.4 The MINERVA audit

In this section we rigorously describe the MINERVA risk-limiting audit. The stopping condition for BRAVO is a comparison test for the ratio of probabilities of the number of winner ballots. On the other hand, the stopping condition for the MINERVA test is a comparison test for the ratio of the complementary cumulative distribution functions (cdfs). For the MINERVA audit, the stopping condition for a given round does depend on previous round sizes, which are required to compute the complementary cdfs, but not on future round sizes. Our proofs for its risk-limiting property assume that the rounds are pre-determined.

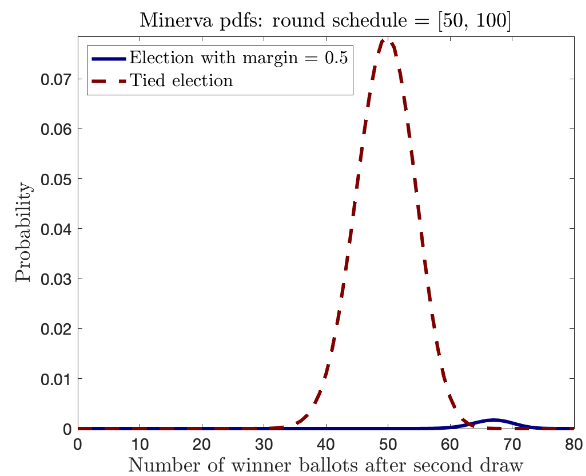

Figure 5: *Probability Distribution of Winner Ballots for* MINERVA*:* $p = 0.75, n_1 = 50, n_2 = 100$*: After Drawing the Second Round.* The distributions are generated by the convolution procedure described. Notice that, if the underlying election were as announced (solid blue line), a large fraction of the samples would satisfy the stopping condition, and a small fraction would proceed to the next round. On the other hand, if the underlying election is tied (dashed red line), the sample is far less likely to satisfy the stopping condition and many audits would proceed to the next round.

Given the B2 (α, p)-BRAVO test we define the corresponding R2 MINERVA test by its stopping condition, which is a comparison test of the ratio of the complementary cdfs of samples that did not satisfy the stopping condition for any previous round.

Definition 3 (($\alpha, p, (n_1, n_2, \ldots, n_j, \ldots)$)-MINERVA)**.** *Given B2* (α, p)-BRAVO *and round sizes* $n_1, n_2, \ldots, n_j \ldots$*, the corresponding R2* MINERVA *stopping rule for the* j^{th} *round is:*

$$\mathcal{A}(X_j) = \begin{cases} Correct & \tau_j(k_j, p, (n_1, \ldots, n_j), \alpha) \geq \frac{1}{\alpha} \\ Undetermined & else \end{cases} \quad (6)$$

where τ_j *is the* complementary cumulative distribution ratio *for the* j^{th} *round, for* $j \geq 2$*:*

$$\tau_j(k_j, p, (n_1, \ldots, n_j), \alpha) = \frac{Pr[K_j \geq k_j \wedge \forall_{i<j}(\mathcal{A}(X_i) \neq Correct) \mid H_a, n_1, \ldots, n_j]}{Pr[K_j \geq k_j \wedge \forall_{i<j}(\mathcal{A}(X_i) \neq Correct) \mid H_0, n_1, \ldots, n_j]} \quad (7)$$

and, as with B2 (α, p)-BRAVO*,* H_a*, the alternate hypothesis, is that the fractional tally for the winner is* p*.*

Clearly, for $j = 1$ and the first round,

$$\tau_1(k_1, p, n_1) = \frac{Pr[K_1 \geq k_1 \mid H_a, n_1]}{Pr[K_1 \geq k_1 \mid H_0, n_1]}.$$

5 MINERVA properties

In this section, we present properties of MINERVA, formulated as theorems. The proofs of technical lemmas are in the Appendix. We begin with notation, then show that MINERVA is risk-limiting and that its B2 version stops exactly when the B2 version of BRAVO does. Finally, we show that MINERVA is at least as efficient as *end-of-round* BRAVO. Before each theorem and proof we provide an informal explanation of the result.

5.1 Notation and Definitions

First, we establish some shorthand notation which will be useful. For ease of notation, when the audit and its parameters (round schedule, risk limit, fractional vote for the winner) are fixed, we denote:

$$S_j(k_j) \triangleq Pr[K_j \geq k_j \wedge \forall_{i<j}(\mathcal{A}(X_i) \neq Correct) \mid H_a, n_1, \ldots, n_j]$$

$$R_j(k_j) \triangleq Pr[K_j \geq k_j \wedge \forall_{i<j}(\mathcal{A}(X_i) \neq Correct) \mid H_0, n_1, \ldots, n_j]$$

Thus $\frac{S_j(k_j)}{R_j(k_j)}$ is the ratio of the complementary cdfs in round j when the number of winner ballots drawn is K_j, and the sequence did not satisfy the stopping condition in a previous round.
Similarly,

$$s_j(k_j) \triangleq Pr[K_j = k_j \wedge \forall_{i<j}(\mathcal{A}(X_i) \neq Correct) \mid H_a, n_1, \ldots, n_j]$$

and

$$r_j(k_j) \triangleq Pr[K_j = k_j \wedge \forall_{i<j}(\mathcal{A}(X_i) \neq Correct) \mid H_0, n_1, \ldots, n_j]$$

and $\frac{s_j(k_j)}{r_j(k_j)}$ is the likelihood ratio of k_j winner ballots in round j when the sequence did not satisfy the stopping condition in a previous round.

Note also the following simple observation:

$$S_j(k_j) = \sum_{k=k_j}^{n_j} s_j(k), \quad R_j(k_j) = \sum_{k=k_j}^{n_j} r_j(k) \tag{8}$$

Recall that, when we do not refer to parameters at all, S_j corresponds to the stopping probability of the j^{th} round and is not a function of the sample drawn, but of the audit.

Definition 4 (S_j). *The probability of stopping in the j^{th} round for audit $\mathcal{A}$ is defined as: $S_j = Pr[(\mathcal{A}(X_j) = Correct) \wedge \forall_{i<j}(\mathcal{A}(X_i) \neq Correct) \mid H_a, n_1, \ldots, n_j]$.*

Definition 5 (R_j). *The risk of the j^{th} round of audit $\mathcal{A}$ is defined as: $R_j = Pr[(\mathcal{A}(X_j) = Correct) \wedge \forall_{i<j}(\mathcal{A}(X_i) \neq Correct) \mid H_0, n_1, \ldots, n_j]$.*

5.2 MINERVA is risk-limiting

In this section we show that MINERVA is risk-limiting. The idea is simple. In the appendix we state and prove lemmas that show that the BRAVO ratio, σ, is monotone increasing as a function of k, because of which the stopping condition $\sigma \geq \alpha^{-1}$ is equivalent to $k \geq k_{min}$ for some k_{min}. We also show that σ, though defined as the ratio of the pdfs without any truncation or convolution, is also the ratio of the pdfs when they are computed assuming previous round sizes and using convolution, as we described in section 4.3. Using the above results and mathematical induction, we show that the MINERVA ratio τ is also monotone increasing with k and can similarly be represented as $k \geq k'_{min}$ for some (other) k'_{min}.

The tail at $K = k'_{min}$ of a distribution (whether representing the election as announced or tied), $Pr[K \geq k'_{min}]$, in a particular round, is the probability that the corresponding election passes the test in that round. Thus the MINERVA condition simply ensures that the risk of a particular round (probability that the audit stops for the tied election) is α times the stopping probability (probability that the audit stops for the election as announced). Summing over all the rounds, because the total stopping probability cannot be greater than 1, the risk cannot be greater than α.

Theorem 1. *If the round schedule is pre-determined (before the audit begins), $(\alpha, H_a, (n_1, \ldots))$-MINERVA is an α-RLA.*

Proof. From Definition 5 and Lemma 1 (formulated and proved in Appendix A), we have

$$\begin{aligned} R_j &= Pr[K_j \geq k_{min,j}(\text{MINERVA}, (n_1, \ldots), p, \alpha) \mid H_0, n_1, \ldots, n_j] \\ &\leq \alpha Pr[K_j \geq k_{min,j}(\text{MINERVA}, (n_1, \ldots), p, \alpha) \mid H_a, n_1, \ldots, n_j] \\ &= \alpha \cdot S_j \end{aligned}$$

because $k_{min,j}(\text{MINERVA}, (n_1, \ldots), p, \alpha)$ satisfies the MINERVA stopping condition.

Define the total stopping probability of the audit as follows: $S = Pr[(\mathcal{A}(X) = Correct) \mid H_a]$.

Then,

$$S = \sum_j S_j \leq 1 \tag{9}$$

The risk of the audit is defined as:
$R = Pr[(\mathcal{A}(X) = Correct) \mid H_0] = \sum_j R_j \leq$
$\leq \alpha \cdot \sum_j S_j = \alpha \cdot S \leq \alpha$ from Equation (9). ☐

5.3 Properties of B2 version of MINERVA

In this section we study the relationship between B2 BRAVO and MINERVA with each round consisting of a single ballot draw. We show that samples satisfying the stopping condition of (α, p)-BRAVO, performed ballot-by-ballot, are exactly those satisfying that of the $(\alpha, H_a, (1, 2, 3, \ldots, j, \ldots))$-MINERVA audit, where H_a is the hypothesis that the winner's fractional tally is p. The p-values of the two audits, however, differ except at their values of k_{min}.

This result follows because, when performed ballot-by-ballot, both audits stop when the number of winner ballots is the largest possible (otherwise the audit would have stopped in the previous round because only one ballot was drawn). For these values, the MINERVA and BRAVO ratios are identical because the tails consist of a single value.

Theorem 2. *The B2* (α, p)-BRAVO *audit stops for a sample of size* n_j *with* k_j *ballots for the winner, if and only if the* $(\alpha, H_a, (1,2,3,\ldots,j,\ldots))$-MINERVA *audit stops.*

Proof. Consider the j^{th} round of the MINERVA audit: the j^{th} ballot draw. Suppose that, before the j^{th} round is drawn, and after the stopping condition is tested for the $(j-1)^{th}$ round and the audit stopped if it is satisfied, k is the largest value of winner ballots possible. It is strictly smaller than the corresponding $k_{min,j-1}$, because the audit has stopped for all other values. Further, because at most one winner ballot will be drawn in the j^{th} round, the largest possible number of winner ballots in the j^{th} round is $k+1$.

More formally, let the largest value of k_{j-1} for which $s^*_{j-1}(k_{j-1}) \neq 0$ be k, where s^*_j is as defined in the proof of Theorem 1. Then $k < k_{min,j-1}(\text{MINERVA}, (n_1,n_2,\ldots,n_{j-1},\ldots), p, \alpha)$ by the definition of $k_{min,j-1}$, Theorem 1. Further, the largest value of k_j for which $s_j(k_j) \neq 0$ is $k+1$.

We now show that if the j^{th} round stops at all, it will be for $k_j = k+1$ and no other values of k_j.

We observe that the only way to obtain $k+1$ ballots in the j^{th} round is if the existing number of winner ballots is k and the new ballot drawn is for the winner. The probability is $s_j(k+1) = ps_{j-1}(k)$. On the other hand, k ballots arise in the j^{th} round if the existing number is $k-1$ and a winner ballot is drawn, or the existing number is k and the ballot drawn is not for the winner. Hence $s_j(k) = (1-p)s_{j-1}(k) + ps_{j-1}(k-1)$. Similarly: $r_j(k+1) = \frac{1}{2}r_{j-1}(k)$ and $r_j(k) = \frac{1}{2}r_{j-1}(k) + \frac{1}{2}r_{j-1}(k-1)$.

If the condition is satisfied by values other than $k+1$, because τ is monotone increasing, it is satisfied by k: $\tau_j(k) = \frac{s_j(k+1)+s_j(k)}{r_j(k+1)+r_j(k)} = \frac{s_{j-1}(k)+ps_{j-1}(k-1)}{r_{j-1}(k)+\frac{1}{2}r_{j-1}(k-1)} \geq \frac{1}{\alpha}$.

Thus $\tau_j(k)$ is a weighted average of $\sigma(k,p,j-1)$ and $\frac{p}{\frac{1}{2}}\sigma(k-1,p,j-1)$ and:

$\frac{p}{\frac{1}{2}}\sigma(k-1,p,j-1) = \frac{(1-p)}{\frac{1}{2}}\sigma(k,p,j-1)$
$< \sigma(k,p,j-1) < \tau(k,p,j-1) < \frac{1}{\alpha}$

as $k < k_{min,j-1}(\text{MINERVA}, (1,2,\ldots,j-1,\ldots), p, \alpha)$. And hence, $\tau_j(k)$ does not pass the stopping condition.

Thus, if $\mathcal{A}_M$ and $\mathcal{A}_B$ denote the B2 MINERVA and B2 BRAVO audits respectively,

$\mathcal{A}_M(X_j) = Correct \Leftrightarrow \tau_j(k,p,j) \geq \frac{1}{\alpha}$
$\Leftrightarrow \sigma(k,p,j) \geq \frac{1}{\alpha} \Leftrightarrow \mathcal{A}_B(X_j) = Correct$.

Samples that do satisfy the stopping condition have the same MINERVA and BRAVO p-values, which are otherwise not the same. □

5.4 Efficiency

In this section we present an efficiency result for MINERVA. The proof is simple, and an intuition for it was developed in Example 4, Section 4.2. The ratio τ at a value $K = k$ is a weighted average of the values of σ for $K \geq k$, and σ is monotonic increasing, thus $\tau \geq \sigma$, with equality occurring when the tail consists of a single value.

Theorem 3. *Given sample X of size* n_j *with* k_j *samples for the winner,* $\mathcal{A}_B(X) = Correct \Rightarrow \mathcal{A}_M(X) = Correct$ *where* $\mathcal{A}_B$ *denotes the* (α, p)-BRAVO *test and* $\mathcal{A}_M$ *the* $(\alpha, p, (n_1,n_2,\ldots,n_j,\ldots))$-MINERVA.

Proof. For a fixed election and fixed round sizes, each of τ and ω is a weighted sum of values of σ, which is monotone increasing, and generally larger than σ. In fact, equality for τ occurs only when k is the largest possible number of winner ballots in the round. Thus
$\sigma(k_j,p,n_j) \geq \frac{1}{\alpha} \Rightarrow \tau_j(k_j,p,(n_1,n_2,\ldots,n_j,\ldots)) \geq \frac{1}{\alpha}$
and $\omega_j(k_j,p,(n_1,n_2,\ldots,n_j,\ldots)) \geq \frac{1}{\alpha}$. □

From Theorem 3 it follows that MINERVA is at least as efficient as the corresponding *end-of-round* application of B2 rules. In Section 6 we demonstrate that MINERVA can be considerably more efficient.

6 Applications

In this section we describe applications of our results. The Appendix contains more verification, and provides some more detail on our work.

Table 1 presents our estimates for the number of distinct ballots in first round sizes for both *end-of-round* BRAVO and MINERVA. These values are computed for 90% stopping probability and a risk limit of 0.1, for the announced statewide results of the 2020 US Presidential election, as obtained from the MIT Election Data and Science Lab [10], for selected states. We constructed a table of stopping probability as a function of round size for a given margin, where the stopping probability of a round is the tail corresponding to the k_{min} value for that round size. We used this to compute an estimate of the round sizes in expected number of distinct ballots drawn, see Appendix Section 2 for details.

It is noteworthy that, across all margins, *end-of-round* BRAVO first round sizes are about twice those of MINERVA: the mean value of the ratio of *end-of-round* BRAVO sizes to MINERVA sizes is 1.9604, and the median is 1.9964. Table 5 in the Appendix presents these round size estimates for all states.

We also estimate first round sizes for 90% stopping probability for *selection-ordered-ballots* BRAVO by treating it as a multiple-round audit. Our results are presented in Table 5 in the Appendix. We notice that among the three types of audits, MINERVA requires the fewest number of ballots and *end of round* BRAVO the largest, for every state.

For the comparison with *selection-ordered-ballots* BRAVO we currently omit estimates for states with margins smaller than 0.025. In the other states, we observe that the increase in round size on using *selection-ordered-ballots* BRAVO over MINERVA is $25\% - 39.5\%$, with greater improvements for smaller margins. The median value of the ratio of *selection-ordered-ballots* BRAVO sizes to Minerva sizes is 1.3507 and the mean is 1.3387. Recall that, unlike *selection-ordered-ballots* BRAVO, MINERVA does not require that the ballots be noted in selection order; sample tallies are sufficient. Thus the reduction in effort due to both—the reduction in total number of ballots, and no longer needing to note ballots in selection order—is considerable.

State	Margin	EoR BRAVO Ballots	MINERVA Ballots
Arizona	0.0031	1,196,732	640,652
Colorado	0.1388	774	384
District of Columbia	0.8893	14	8
Florida	0.0339	12,530	6,070
Michigan	0.0283	18,161	8,807
Nevada	0.0245	24,311	11,783
North Carolina	0.0137	76,857	37,303
Pennsylvania	0.0118	103,559	50,092
Texas	0.0566	4,520	2,221

Table 1: Comparison of *end-of-round* (EoR) BRAVO and MINERVA First-Round Sizes (in distinct ballots) for Statewide 2020 US Presidential Contests, for a stopping probability of 0.9, for selected states.

Of course, some of these sizes are too large for consideration in a real audit. For example, the effort associated with drawing the fraction 0.2170 of ballots for a MINERVA ballot polling audit in Georgia would likely be much larger than that of counting all the ballots, and the effort for both BRAVO audits would be even larger.

Figure 6 plots the *end-of-round* BRAVO and *selection-ordered-ballots* BRAVO round sizes as a fraction of the corresponding MINERVA round size. There is a small variation with margin, with the fraction being larger for smaller margins. Note that a couple of states with the smallest margins do not have the largest ratios for *end-of-round* BRAVO. For these states, the ratios of random draws are among the largest, but the number of random draws for *end-of-round* BRAVO are such a large fraction of the total that considering distinct ballots instead of random draws changes the ratio. For example, in Georgia, MINERVA requires that the number of random draws be 0.2446 of the total, and *end-of-round* BRAVO requires 0.5086, for a ratio of 2.08. Considering distinct ballots draws reduces the ratio to 1.84.

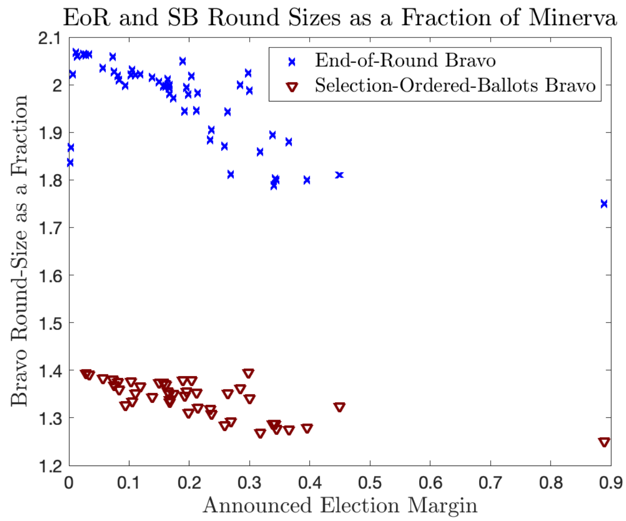

Figure 6: Ratios: MINERVA first-round sizes in expected number of distinct ballots drawn for 90% stopping probability as a fraction of those of *End-of-Round* BRAVO and *Selection-Ordered-Ballots* BRAVO, for the statewide margins of the 2020 US Presidential contest.

7 Montgomery County OH audit

MINERVA was used by Mark Lindeman of Verified Voting for a pilot audit of the 2020 primary elections in Montgomery County, Ohio. There were a number of contests on the ballot [11], of which three were audited: the Democratic Presidential primary (10 candidates), the Democratic County Commissioner FTC 1-2-2021 (2 candidates) and the Republican County Commissioner FTC 1-2-2021 (2 candidates). Ballots were not stored separately based on party, but the contests were, by definition, partisan. A total of 69,743 ballots were cast.

We estimate the number of ballots for the first round, for a given probability of stopping, for the closest contest—that for the Republican County Commissioner FTC 1-2-2021. It had 15691 votes for Candidate Setzer and 8538 for Candidate Scearce. All ballots in the sample that do not bear a vote for either of the candidates are ignored. The relevant margin is thus (15691-8538)/(15691+8538) $\approx$ 0.2952, which yields an estimate of 79 ballots to be picked for a 90% stopping probability. We scale that up by a factor of 2.88 to account for the fact that only one in 2.88 ballots is relevant for this contest, yielding an estimate of 228. For the actual audit, 240 ballots were selected for the first round for convenience. Table 2 presents the predicted round sizes for stopping probabilities of 0.7, 0.8 and 0.9.

Audit	Round Sizes			$\alpha =$	$\alpha =$	Final
	0.7	0.8	0.9	0.1	0.05	p-value
MINERVA	150	179	228	89	94	0.0019
SB	176	231	326	92	101	0.0022
EoR	251	340	475	92	101	0.0034

Table 2: Performance of MINERVA, *selection-ordered-ballots* BRAVO (SB) and *end-of-round* BRAVO (EoR) on the data from the audit of the 2020 primaries of Montgomery County, Ohio, for the County Commissioner, FTC 1-2-2021 (R) contest, with a margin of 0.2952. Estimated round sizes for various stopping probabilities for $\alpha = 0.1$ are noted, as are the smallest sizes of the actual sample that would have been sufficient in this audit for $\alpha = 0.1$ and $\alpha = 0.05$. Also noted are the final p-values. Observe that estimated round sizes are smallest for MINERVA, the MINERVA audit would have ended earliest, and the MINERVA p-value is smallest, as expected.

A ballot manifest was prepared, which assigns to each ballot a unique identifier based on where it is stored (the 100^{th} ballot in the 5^{th} box, for example). A 20-digit random number from dice rolls was fed in to Arlo's implementation of consistent sampling [17], yielding an ordered sample of 240 ballot ids, unpredictable in advance. The ballot ids were sorted by box and id, and the sample ballots were pulled and manually examined. The votes on each ballot were recorded on tally sheets. The sample tallies for Setzer and Scearce were 49 and 20, yielding a MINERVA p-value of 0.0019, which was below the risk limit. The other contests also met their risk limits, so the audit could end, having achieved its goal. It was a relatively lucky draw. Both other methods also met the risk limit, but only MINERVA could predict success (assuming accurate results) with 90% probability.

In order to illustrate how each audit would work with different round sizes, we retrospectively used the tally sheet data and the original selection ordering to calculate the risk levels that would have been computed by MINERVA, selection-ordered BRAVO (SB) and end-of-round BRAVO (EoR) for round sizes 1 through 240. That is also how we calculated the minimum number of draws at which the audit would achieve the risk limit given the actual roll of the dice, as shown in Table 2.

We have plotted the p-value as a function of number of ballots drawn, for this particular sample of ballots. Figure 7 shows the p-values as a function of number of ballots for the three audits.

The close-up of the corresponding plot for the County Commissioner, FTC 1-2-2021 (D) contest, with a margin of 0.3833, in Figure 8, further illustrates the variation in p-values.

We see that MINERVA provides an advantage in this instance as well, see Table 3.

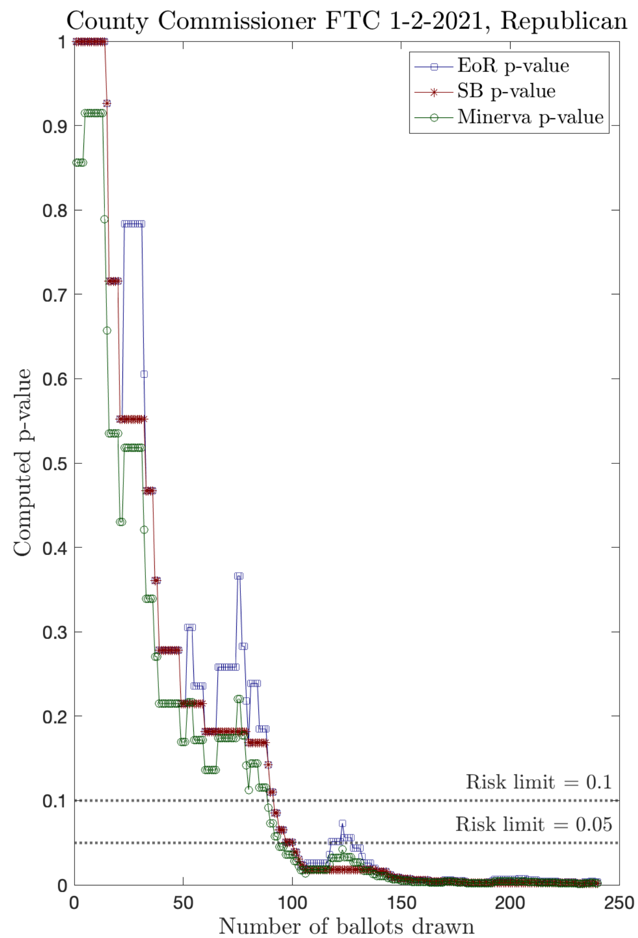

Figure 7: p-values for MINERVA, *End-of-Round* BRAVO (EoR) and *Selection-Ordered-Ballots* BRAVO (SB), as a function of the number of ballots drawn for the audit of County Commissioner, FTC 1-2-2021 (R) contest in the 2020 primaries in Montgomery County, Ohio. Notice that MINERVA has the lowest p-value except for a few values around 125 ballots, and that EoR always has the largest p-value, as expected. Recall that the SB p-value is, by definition, the smallest p-value of all ballot draws so far. Notice that it, hence, does not increase. Also notice that there are many instances when the p-values do not change, because the next ballot picked was not voted for either candidate in the closest pair.

8 Conclusion

We describe inefficiencies with the use of audits developed for ballot-by-ballot decisions in round-to-round procedures, such as are in use in real audits today. We propose a new audit, MINERVA, which we prove is risk-limiting if the round sizes are pre-determined, and at least as efficient as audits that apply the ballot-by-ballot decision rules at the end of the round.

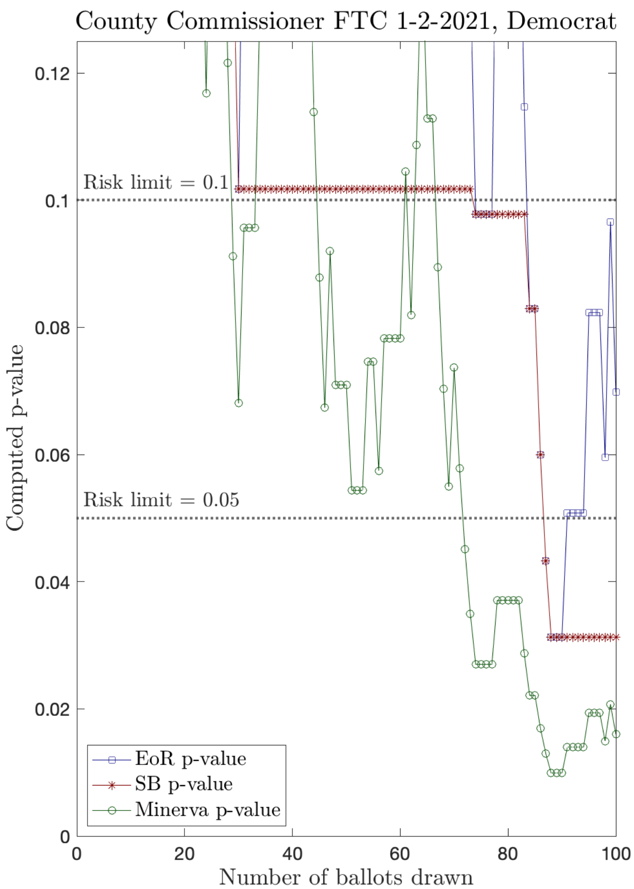

Figure 8: p-values for MINERVA, *End-of-Round* BRAVO (EoR) and *Selection-Ordered-Ballots* BRAVO (SB), as a function of the number of ballots drawn for the audit of County Commissioner, FTC 1-2-2021 (D) contest in the 2020 primaries in Montgomery County, Ohio. Notice that MINERVA p-value is not always the smallest, and that EoR p-value is always largest. Observe that the SB p-value takes on the value 0.1 for the first time after the MINERVA p-value goes below 0.05 for the first time.

We describe an approach to computing stopping probabilities and risks of audits with stopping conditions that are monotone increasing with the number of ballots for the winner in the sample. We demonstrate its accuracy in reproducing the empirically-obtained percentile values from [9, Table 1]: the average absolute fractional discrepancy is just 0.13%.

We predict first round sizes (for 90% stopping probability) for representative states in the US Presidential election of 2020 for *end-of-round* BRAVO and MINERVA. We find that our proposed audits require half the ballots for the commonly-used 90% stopping probability across all margins. We similarly compare first round sizes to *selection-ordered-ballots* BRAVO as well, finding that it requires 25%-39.5% more ballots than does MINERVA, with the larger improvements corresponding to smaller margins. We thus see that the additional effort of retaining information by ballot id, required by *selection-ordered-ballots* BRAVO, is not beneficial as the MINERVA class of audits does not require it.

Audit	Round Sizes			$\alpha =$	$\alpha =$	Final
	0.7	0.8	0.9	0.1	0.05	p-value
MINERVA	68	76	101	29	72	0.00089
SB	70	93	129	74	87	0.00016
EoR	99	129	188	74	87	0.0146

Table 3: Performance of MINERVA, SB and EoR on the data from the audit of the 2020 primaries of Montgomery County, Ohio, for the County Commissioner, FTC 1-2-2021 (D) contest, with a margin of 0.3833. Estimated round sizes for various stopping probabilities for $\alpha = 0.1$ are noted, as are the smallest sizes of the sample that would have been sufficient in this audit for $\alpha = 0.1$ and $\alpha = 0.05$. Also noted are the final p-values. Observe that estimated round sizes are smallest for MINERVA and that the MINERVA audit would have ended earliest. The final MINERVA p-value is larger than the SB p-value, as would be expected to happen on occasion.

MINERVA was used in May 2020 for a pilot RLA of the primaries in Montgomery County, Ohio.

We provide open-source software for computing probability distributions and for the MINERVA audit, hoping it helps developers of election auditing software. We also hope our work, including the code integrated as an option into the most popular election audit software package Arlo, helps more election officials implement more audits, more efficiently.

9 Acknowledgements

This research was supported in part by NSF Awards 2015253 and 1421373 and Polish National Science Centre contract number DEC-2013/09/D/ST6/03927. The data on the Montgomery County primary of 2020 was collected by Mark Lindeman who carried out the audit; we are very grateful for his early adoption of MINERVA and his encouragement. The authors gratefully acknowledge careful readings of an earlier draft and valuable suggestions by Matthew Bernhard, Amanda Glazer, Mark Lindeman, Jake Spertus, Mayuri Sridhar, Philip B. Stark and Damjan Vukcevic. In particular, Philip B. Stark pointed out that our proofs are valid only when round sizes are pre-determined. Discussions with Oliver Broadrick and Ronald L. Rivest were very helpful. Ruth Godberfforde helped organize the data from the Montgomery County, Ohio primary.

References

[1] Matthew Bernhard, Allison McDonald, Henry Meng, Jensen Hwa, Nakul Bajaj, Kevin Chang, and J Alex Halderman. Can voters detect malicious manipulation of ballot marking devices? In *41st IEEE Symposium on Security and Privacy*, 2020.

[2] Joseph A Calandrino, Ariel J Feldman, J Alex Halderman, David Wagner, Harlan Yu, and William P Zeller. Source code review of the Diebold voting system. *University of California, Berkeley under contract to the California Secretary of State*, 2007.

[3] Stephen Checkoway, Ariel J Feldman, Brian Kantor, J Alex Halderman, Edward W Felten, and Hovav Shacham. Can DREs provide long-lasting security? the case of return-oriented programming and the AVC advantage. *EVT/WOTE*, 2009, 2009.

[4] Ariel J Feldman, J Alex Halderman, and Edward W Felten. Security analysis of the Diebold AccuVote-TS voting machine, 2006.

[5] Lloyd D Fisher. Self-designing clinical trials. *Statistics in medicine*, 17(14):1551–1562, 1998.

[6] Bhaskar Kumar Ghosh and Pranab Kumar Sen. *Handbook of sequential analysis*. CRC Press, 1991.

[7] Nicholas A Heard and Patrick Rubin-Delanchy. Choosing between methods of combining-values. *Biometrika*, 105(1):239–246, 2018.

[8] Mark Lindeman and Philip B Stark. A gentle introduction to risk-limiting audits. *IEEE Security & Privacy*, 10(5):42–49, 2012.

[9] Mark Lindeman, Philip B Stark, and Vincent S Yates. BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In *EVT/WOTE*, 2012.

[10] MIT Election Data and Science Lab. U. S. President, 1976-2020, https://electionlab.mit.edu/data.

[11] Ohio Montgomery County. Election summary report official primary election march 17, 2020. https://www.montgomery.boe.ohio.gov/download/336/2020/9561/2020-march-17-2020-primary-election-summary.pdf.

[12] Sarah Morin and Grant McClearn. The R2B2 (Round-by-Round, Ballot-by-Ballot) library, https://github.com/gwexploratoryaudits/r2b2.

[13] National Academies of Sciences, Engineering, and Medicine. *Securing the Vote: Protecting American Democracy*. The National Academies Press, Washington, DC, 2018.

[14] Kellie Ottoboni, Philip B. Stark, Mark Lindeman, and Neal McBurnett. Risk-limiting audits by stratified union-intersection tests of elections (SUITE). In *E-Vote-ID 2018*, pages 174–188, 2018.

[15] Ronald L. Rivest. Clipaudit—a simple post-election risk-limiting audit. arXiv:1701.08312.

[16] Ronald L Rivest. On the notion of "software independence" in voting systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1881):3759–3767, 2008.

[17] Ronald L. Rivest. Consistent sampling with replacement. *CoRR*, abs/1808.10016, 2018.

[18] Mayuri Sridhar and Ronald L. Rivest. k-Cut: A simple approximately-uniform method for sampling ballots in post-election audits. In Andrea Bracciali, Jeremy Clark, Federico Pintore, Peter B. Rønne, and Massimiliano Sala, editors, *Financial Cryptography and Data Security - FC 2019 International Workshops, VOTING and WTSC, St. Kitts, St. Kitts and Nevis, February 18-22, 2019, Revised Selected Papers*, volume 11599 of *Lecture Notes in Computer Science*, pages 242–256. Springer, 2019.

[19] Philip B. Stark. Personal communication.

[20] Philip B. Stark. Conservative statistical post-election audits. *Ann. Appl. Stat.*, 2(2):550–581, 06 2008.

[21] Philip B. Stark and David A. Wagner. Evidence-based elections. *IEEE Secur. Priv.*, 10(5):33–41, 2012.

[22] Technical Guidelines Development Committee. Voluntary Voting System Guidelines Version 2.0 (Draft, Clean Version), 2020.

[23] Poorvi L. Vora. brla_explore, https://github.com/gwexploratoryaudits/brla_explore.

[24] Poorvi L. Vora. Risk-limiting Bayesian polling audits for two candidate elections. *CoRR*, abs/1902.00999, 2019.

[25] VotingWorks. Arlo, https://voting.works/risk-limiting-audits/.

[26] Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

[27] Gernot Wassmer. Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers*, 41(3):253–279, 2000.

[28] Filip Zagórski. Athena - risk limiting audit (round-by-round), https://github.com/filipzz/athena.

A Proofs

In this section we state interesting properties of the MINERVA ratio.

The B2 (α, p)-BRAVO stopping condition is a test of the ratio $\sigma(k,p,n)$. The history of round size is completely captured in the total number of ballots drawn, n, the ratio tested is monotone increasing with k, and hence the test is a comparison test for k. We show that $\sigma(k,p,n)$ is also the likelihood ratio of winner ballots in all rounds of the MINERVA audit, even though round sizes are not constrained in any way. Additionally, we show that the ratio tested for the MINERVA audit are also monotone increasing, and hence that the tests are also comparison tests for k.

Lemma 1. *For the $(\alpha, p, (n_1, n_2, \ldots, n_j, \ldots))$-MINERVA test, the following are true for $j = 1, 2, 3, \ldots$*

1. $$\frac{s_j(k_j)}{r_j(k_j)} = \sigma(k_j, p, n_j)$$
   when $r_j(k_j)$ and $s_j(k_j)$ are defined and non-zero.

2. $\tau_j(k_j, p, (n_1, n_2, \ldots, n_j), \alpha)$ *is monotone increasing as a function of k_j.*

3. $\exists k_{min,j}(\text{MINERVA}, (n_1, n_2, \ldots, n_j, \ldots), p, \alpha)$ *such that*
   $$\mathcal{A}(X_j) = Correct$$
   $$\Leftrightarrow k_j \geq k_{min,j}(\text{MINERVA}, (n_1, n_2, \ldots, n_j, \ldots), p, \alpha)$$

We need the following general results from basic algebra.

Lemma 2. *Given a monotone increasing sequence: $\frac{a_1}{b_1}, \frac{a_2}{b_2}, \ldots, \frac{a_n}{b_n}$, for $a_i, b_i > 0$, the sequence: $z_i = \frac{\sum_{j=i}^n a_j}{\sum_{j=i}^n b_j}$ is also monotone increasing.*

Proof. Note that z_i is a weighted average of the values of $\frac{a_j}{b_j}$ for $j \geq i$: $z_i = \sum_{j=i}^n y_j \frac{a_j}{b_j}$ for $y_j = \frac{b_j}{\sum_{j=i}^n b_j} > 0$. Further, $\sum_{j=i}^n y_j = 1$ and hence $y_j \leq 1$ and $y_j = 1 \Leftrightarrow i = j = n$. Observe that, because $\frac{a_i}{b_i}$ is monotone increasing, $z_i \geq \frac{a_i}{b_i}$ with equality if and only if $i = n$. Suppose $i < n$. Then $z_{i+1} \geq \frac{a_{i+1}}{b_{i+1}} > \frac{a_i}{b_i}$, and $z_i = y_i \frac{a_i}{b_i} + (1 - y_i) z_{i+1} < z_{i+1}$. Thus z_i is also monotone increasing. □

Lemma 3. *Given a strictly monotone increasing sequence: $x_1, x_2, \ldots x_n$ and some constant A, $\exists i_{min}$ such that $x_i \geq A \Leftrightarrow i \geq i_{min}$.*

Proof. Clear. □

Lemma 4. *Given p, n, with $p > \frac{1}{2}$, $\sigma(k,p,n)$ is strictly monotone increasing as a function of k.*

Proof. $p > \frac{1}{2} \Rightarrow p > 1 - p \Rightarrow \frac{1-p}{p} < 1$. Let $0 \leq k < n$. Then: $\sigma(k,p,n) = \frac{1-p}{p} \sigma(k+1,p,n) < \sigma(k+1,p,n)$. □

Now, we are ready to prove Lemma 1.

Proof. We show this by induction.

Consider $j = 1$.

1. $\frac{s_1(k_1)}{r_1(k_1)} = \frac{Pr[K_1 = k_1 | H_a, n_1]}{Pr[K_1 = k_1 | H_0, n_1]} = \sigma(k_1, p, n_1)$.

2. $\tau_1(k_1, p, n_1) = \frac{Pr[K_1 \geq k_1 | H_a, n_1]}{Pr[K_1 \geq k_1 | H_0, n_1]} = \frac{S_1(k_1)}{R_1(k_1)} = \frac{\sum_{k=k_1}^{k_{max,1}} s_1(k)}{\sum_{k_{max,1}}^{n} r_1(k)}$, where $k_{max,j}$ is the largest possible value for k_j. Note that $k_{max,1} = n_1$. $\tau_1(k_1, p, n_1)$ is a weighted average of $\sigma(k, p, n_1)$ for $k \geq k_1$, and, by Lemmas 2 and 4, is strictly monotone increasing as a function of k_1.

3. From Lemma 3, $\exists k_{min,1}(\text{MINERVA}, (n_1, \ldots, n_r, \ldots), p, \alpha)$ such that $\tau_1(k_1, p, n_1) \geq \frac{1}{\alpha} \Leftrightarrow k_1 \geq k_{min,1}(\text{MINERVA}, (n_1, n_2, \ldots, n_r, \ldots), p, \alpha)$, which is the Minerva stopping condition.

Thus the theorem is true for $j = 1$.

Suppose the theorem is true for $j = m$. We will show it is true for $j = m + 1$.

From property (3) of this theorem for $j = m$, we observe that, after the stopping decision is made and before the next round is drawn, the number of winner ballots in the sample is strictly smaller than $k_{min,m}(\text{MINERVA}, (n_1, n_2, \ldots, n_m, \ldots), p, \alpha)$. The distribution on the winner votes may be modeled as $s_m^*(k_m)$ and $r_m^*(k_m)$ where:

$$s_m^*(k_m) = \begin{cases} s_m(k_m) & k < k_{min,m}(\text{MINERVA}, n, p, \alpha) \\ 0 & else \end{cases}$$

where, for space reasons, $k_{min,m}(\text{MINERVA}, n, p, \alpha)$ represents $k_{min,m}(\text{MINERVA}, (n_1, n_2, \ldots, n_r, \ldots), p, \alpha)$ and

$$r_m^*(k_m) = \begin{cases} r_m(k_m) & k < k_{min,m}(\text{MINERVA}, n, \frac{1}{2}, \alpha) \\ 0 & else \end{cases}$$

When we draw the next round of ballots with replacement, the resulting distributions on the winner ballots are convolutions: $s_{m+1} = s_m^* \circledast Bin(p, n_{m+1} - n_m)$ and $r_{m+1} = r_m^* \circledast Bin(0.5, n_{m+1} - n_m)$, where $Bin(p,n)$ represents the binomial distribution for winner ballots in a sample of size n from a distribution with fractional tally p for the winner. Using property (1) of this theorem for $j = m$, we see that $s_m^*(k_m) = A(k_m) p^{k_m} (1-p)^{n_m - k_m}$ and $r_m^*(k_m) = A(k_m) (\frac{1}{2})^{n_m}$ for some A, a function of k_m, current and previous round sizes, p and α.

Some bookkeeping demonstrates that

$$s_{m+1}(k_{m+1}) = B(k_{m+1}) p^{k_{m+1}} (1-p)^{n_{m+1} - k_{m+1}}$$

where $B(k_{m+1}) = A(k_m) \circledast \binom{n_{m+1} - n_m}{k_{new,m+1}}$ and $r_{m+1}(k_{m+1}) = B(k_{m+1}) (\frac{1}{2})^{n_{m+1}}$ which proves property (1) for $j = m + 1$.

Properties (2) and (3) follow for $j = m+1$ by application of Lemmas 2-4.

Thus the theorem is true for all $j \geq 1$.

□

B Experimental Results

B.1 B2 BRAVO Percentile Verification

In this section, we present analytical results for percentiles of the BRAVO stopping condition, and compare them with those reported by Lindeman *et al.* [9, Table 1]. We find that the average absolute value of fractional difference is 0.12%.

We used the approach described in Section 4.3 to generate the probability distributions for B2 BRAVO using various election margins to see how our estimates compared to those obtained by Lindeman *et al.* [9, Table 1]. They used 10,000 simulations.

Table 4 presents our values. Values in parentheses are from [9, Table 1], where they differ. Also listed in the table is Average Sample Number (ASN), which is computed using a standard theoretical estimate (and not using our analytical expressions, nor simulations). It provides a baseline to compare with the values for the Expected Ballots column. Some of the difference between our values and those of [9, Table 1] is likely due to rounding off. Further, we notice that both our values and those of [9, Table 1], when they differ from ASN, are lower than ASN. In our case, the difference is likely due to the fact that we compute our probability distributions for only up to 6ASN draws, using a finite summation to estimate the probability distributions, and we model the discrete character of the problem, which is not captured by ASN. The largest difference between our values and those of [9, Table 1] is 190 ballots, corresponding to a fractional difference of 0.41 %, in the estimate of the expected number of ballots drawn for a margin of 1%. Our value is further from ASN. The average of the absolute value of the fractional difference between our results and those of [9] is 0.13%. The differences between our values and those obtained with simulations could be because 10,000 simulations may not be sufficiently accurate at the lower margins, where most of the errors are. It could also be because our finite summation is not sufficient at low margin.

Margin	25^{th}	50^{th}	75^{th}	90^{th}	99^{th}	Expected Ballots	ASN
0.4	12	22	38	60	131	29.48 (30)	30.03
0.3	23	38	66	108	236	52.85 (53)	53.25
0.2	49	84	149	244	538	118.04 (119)	118.88
0.18	77	131	231	381	842 (840)	183.64 (184)	184.89
0.1	193	332	587	974	2,155 (2,157)	466.55 (469)	469.26
0.08	301	518	916	1,520	3,366	727.04 (730)	730.80
0.06	531	914	1,621 (1,619)	2,698 (2,700)	5,976 (5,980)	1,287.73 (1,294)	1,294.62
0.04	1,190 (1,188)	2,051	3,637	6,055 (6,053)	13,433 (13,455)	2,887.47 (2,900)	2,901.97
0.02	4,727 (4,725)	8,161 (8,157)	14,493 (14,486)	24,155 (24,149)	53,646 (53,640)	11,506.84 (11,556)	11,561.66
0.01	18,845 (18,839)	32,566 (32,547)	57,856 (57,838)	96,469 (96,411)	214,385 (214,491)	45,935.85 (46,126)	46,150.44

Table 4: Computed Estimates of B2 BRAVO Stopping Probability Percentiles. Values in parentheses are those from [9, Table 1] that differ.

B.2 Determining First Round Sizes

For *selection-ordered* BRAVO we use the approach described in Section 4.3 to compute probability distributions for margins 0.025 and above and to find the required number of ballots for the given percentile.

For both *end-of-round* BRAVO and MINERVA we constructed a table of stopping probability as a function of round size for a given margin, where the stopping probability of a round is the tail corresponding to the k_{min} value for that round size. We observed that the stopping probability is not a monotone increasing function of round size. This is because, if k_{min} increases with round size (it does not decrease, but it may remain the same), the stopping probability may decrease slightly.

For both *end-of-round* BRAVO and MINERVA, the round size with a desired stopping probability increases quadratically: reciprocal squared of the margin. As auditors require round size recommendations in real time, a linear search for such round sizes is intractable for tight races. We present here a modified binary search used to compute the round size estimates in this paper. A standard binary search is insufficient because stopping probability is not monotone in round size; larger draws are occasionally marginally less likely to stop than smaller draws.

We conduct modified binary searches on graduating intervals to account for the small round sizes needed for most margins in practice. We emphasize that the difficulty is not in determining the stopping probability of a given round size (that is a straightforward tail computation); it is in the inverse problem.

Ignoring the case when an acceptable round size is less than the absolute lower bound r_0 (which is avoided in practice by selecting r_0 as the lowest possible round size), a round size produced s by this algorithm satisfies the constraint that $s-1$ does not achieve the desired stopping probability. As many round sizes achieve a given stopping probability, this constraint ensures we report a small round size so that auditors do not examine more ballots than necessary.

Note that SPROB(m) computes the stopping probability of a given round size.

In all cases, once we determined the number of ballots required in the contest between the two leading candidates,

Algorithm 1 Modified Binary Search

```
Decide on a desired stopping probability p.
Generate an absolute lower bound, r_0, for the search.
Generate intermediate upper bounds, r_1, r_2, ..., r_n, for graduating searches.
i, s ← 0
while s = 0 and i < n do
    s ← SEARCH(r_i, r_{i+1})
    i ← i + 1
end while
return s
function SEARCH(l, u)
    m ← ⌊(l + u)/2⌋
    if u − l ≤ 1 then
        if SPROB(m) ≥ p then
            return m
        else
            if SPROB(m + 1) ≥ p then
                return m + 1
            end if
        end if
        return 0
    end if
    if SPROB(m) ≥ p then
        return SEARCH(l, m)
    else
        return SEARCH(m, u)
    end if
end function
```

Biden and Trump, we scaled the round size estimates by the ratio of total ballots cast to the number of valid ballots in the contest cast for either Biden or Trump. Finally, we computed the expected number of distinct ballots [19]. We used the approach for states with margin larger than 0.01. For the three states with smaller margin (Arizona, Georgia, Wisconsin) we approximated round size by estimating the binomial as a gaussian.

The table below compares *end-of-round* (EoR) BRAVO, *selection-ordered-ballots* (SB) BRAVO and MINERVA First-Round Sizes (in distinct ballots) for Statewide 2020 US Presidential Contests, for a stopping probability of 0.9.

State	Margin	EoR BRAVO Ballots	SB BRAVO Ballots	MINERVA
Alabama	0.2582	217	149	116
Alaska	0.1052	1359	893	669
Arizona	0.0031	1,196,732	-	640,652
Arkansas	0.2842	182	124	91
California	0.2982	164	113	81
Colorado	0.1388	774	516	384
Connecticut	0.2039	351	240	174
Delaware	0.1925	387	268	199
DistrictOfColumbia	0.8893	14	10	8
Florida	0.0339	12,530	8,442	6,070
Georgia	0.0024	1,993,171	-	1,084,953
Hawaii	0.3007	163	110	82
Idaho	0.3175	145	99	78
Illinois	0.1732	483	331	245
Indiana	0.1639	549	370	273
Iowa	0.0837	2,084	1,410	1,037
Kansas	0.1499	644	441	321
Kentucky	0.2640	204	142	105
Louisiana	0.1893	410	276	200
Maine	0.0934	1,706	1,133	854
Maryland	0.3406	118	85	66
Massachusetts	0.3423	185	66	
Michigan	0.0283	18,161	12,279	8,807
Minnesota	0.0728	2,779	1,864	1,350
Mississippi	0.1677	526	352	263
Missouri	0.1567	587	404	294
Montana	0.1679	523	352	264
Nebraska	0.1957	381	259	191
Nevada	0.0245	24,311	-	11,783
NewHampshire	0.0750	2,600	1,757	1,283
NewJersey	0.1614	555	381	278
NewMexico	0.1104	1,212	811	600
NewYork	0.2343	260	182	138
NorthCarolina	0.0137	76,857	-	37,303
NorthDakota	0.3443	117	83	65
Ohio	0.0815	2,181	1,485	1,080
Oklahoma	0.3388	125	85	66
Oregon	0.1661	534	362	268
Pennsylvania	0.0118	103,559	-	50,092
RhodeIsland	0.2120	319	222	164
SouthCarolina	0.1185	1,043	705	516
SouthDakota	0.2687	192	137	106
Tennessee	0.2366	259	178	136
Texas	0.0566	4,520	3,071	2,221
Utah	0.2139	327	218	165
Vermont	0.3660	109	74	58
Virginia	0.1031	1,368	932	677
Washington	0.1985	382	253	193
WestVirginia	0.3960	90	64	50
Wisconsin	0.0064	338,586	-	167,438
Wyoming	0.4496	67	49	37

Table 5: Comparison of Estimated Round Sizes